Logistic Regression

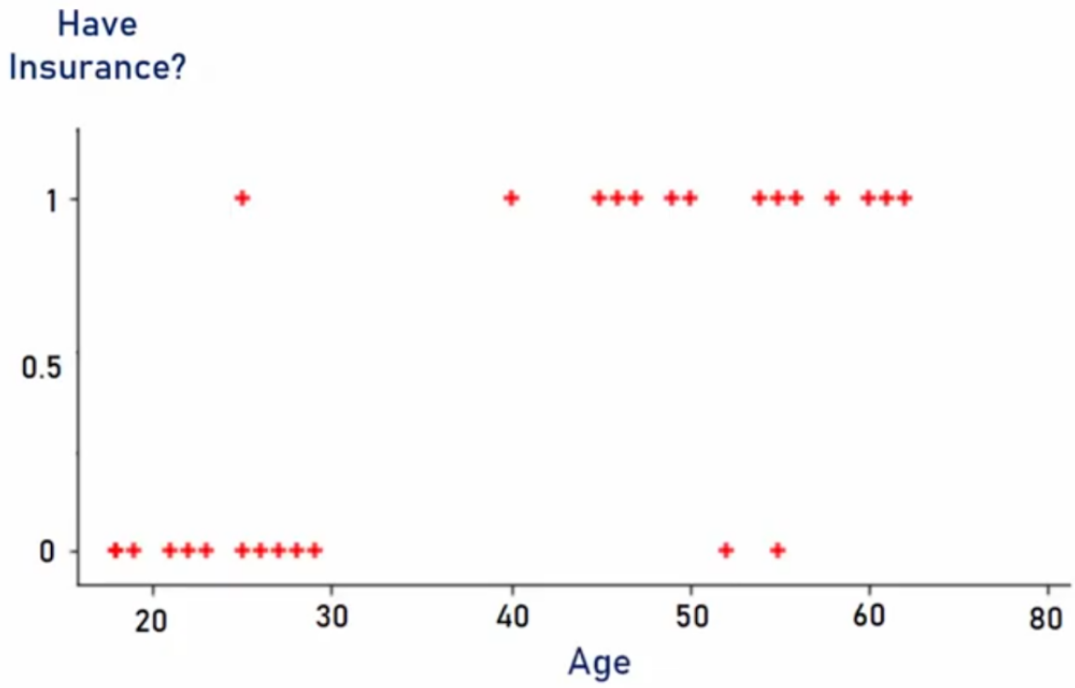
- Logistic regression outputs probabilities
- If the probability ‘p’ is greater than 0.5, The data is labeled ‘1’
- If the probability ‘p’ is less than 0.5 , The data is labeled ‘0’
- By default, logistic regression threshold = 0.5

Logistic Regression Vs Linear Regression

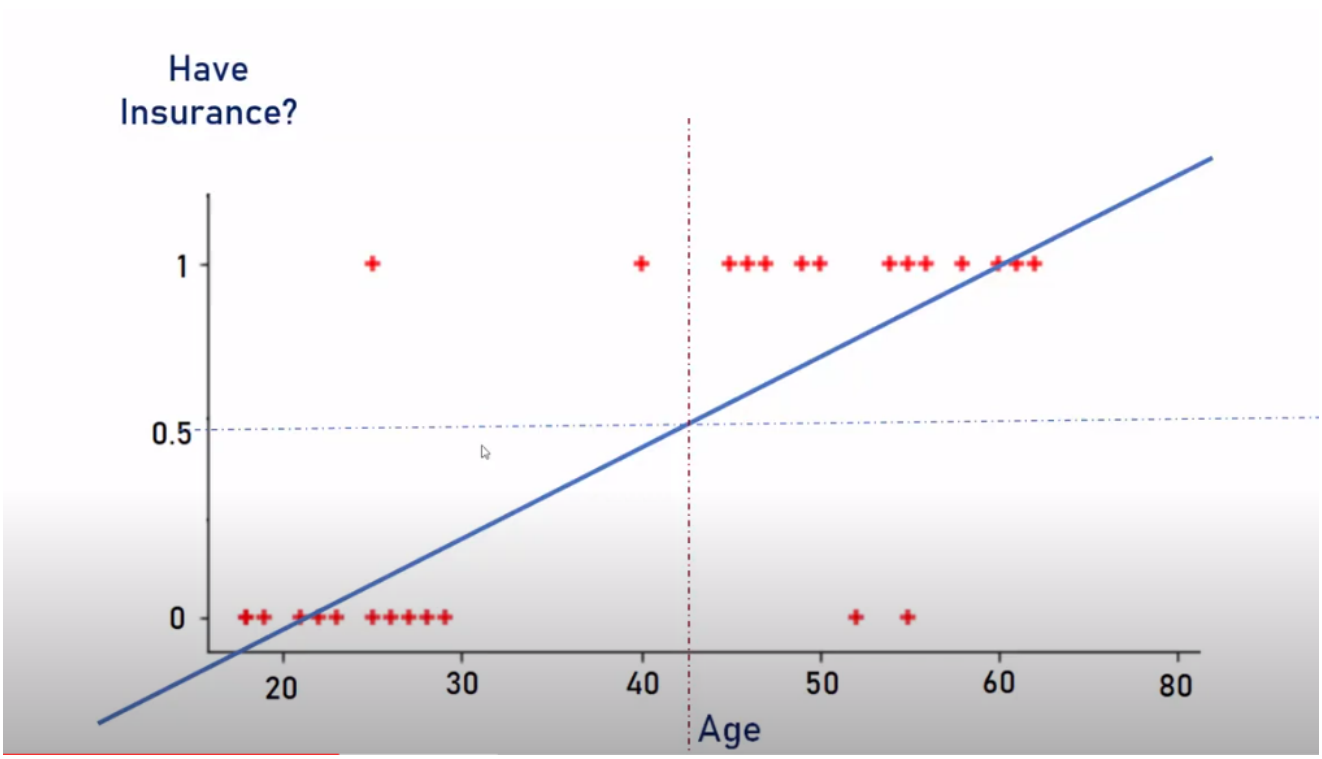
Linear regression is a regression algo which uses OLS to determine the coeff and hence the target var.

- Consider the below scenario, where we are determining people will have insurance or not based on their age using Linear Regression :

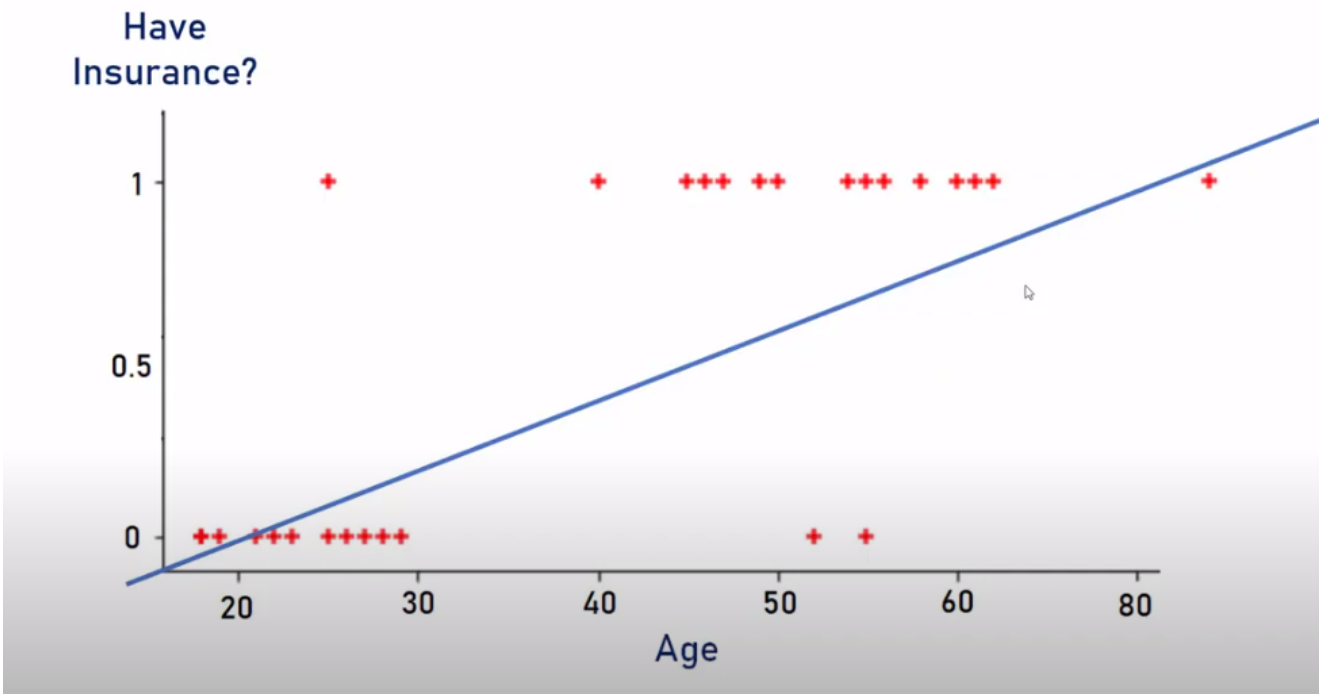
A simple binary data visulization :



Using Linear regression :



Now if we get new person/outlier with higher age opting for Insurance, the OLS will shift the line as :



This will lead to wrong predictions

Sigmoid function

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

e = Euler's number ~ 2.71828

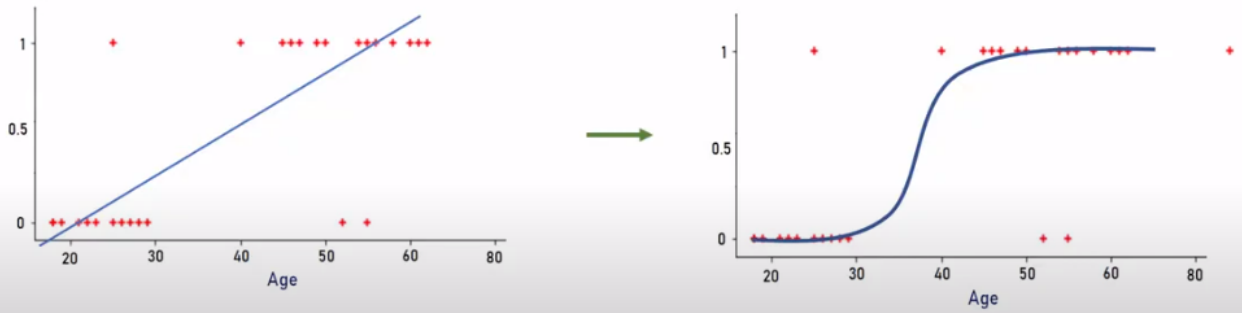
Sigmoid function converts input into range 0 to 1

Hence we convert each data point using sigmoid function to convert the linear graph to the curve

Conversion :

$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$



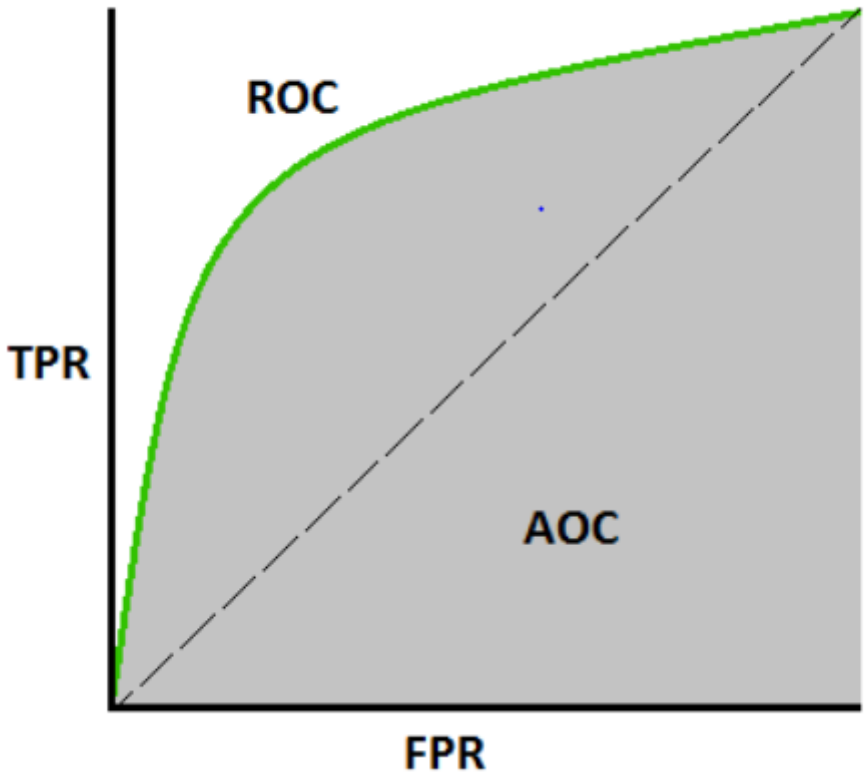
Now the outlier doesn't effect the accuracy of the model as curve covers the necessary data points

- Also with tranformation to sigmoid curve, logistic regression uses concept of MAXIMUM LIKELIHOOD to find the best fit curve just like Linear regression uses OLS
- You calculate Likelihood of every person opting for insurance based on age, and multiply all likelihoods to obtain final likelihood
- Shifts the curve and obtain likelihood for all the the data ponts
- Curve having max final likelihood will be selected.

Link : https://medium.com/30-days-of-machine-learning/day-4-logistic-regression-df9a7a2220cd (https://medium.com/30-days-of-machine-learning/day-4-logistic-regression-df9a7a2220cd)

Performance Measures :

- *AUC - ROC* curve is a performance measurement for classification problem
- It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.
- Larger area under the ROC curve = better model
- The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.



What is **TPR and FPR** ?

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Once you understand these four parameters then we can calculate Accuracy, Precision, Recall and F1 score.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

A confusion matrix is a table that is used to evaluate the performance of a classification model. You can also visualize the performance of an algorithm. The fundamental of a confusion matrix is the number of correct and incorrect predictions are summed up class-wise.

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

In [ ]: